



A generative adversarial network with adaptive constraints for multi-focus image fusion

Jun Huang¹ · Zhuliang Le¹ · Yong Ma¹ · Xiaoguang Mei¹ · Fan Fan¹

Received: 17 January 2020 / Accepted: 14 March 2020 / Published online: 30 March 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

In this paper, we propose a novel end-to-end model for multi-focus image fusion based on generative adversarial networks, termed as ACGAN. In our model, due to the different gradient distribution between the corresponding pixels of two source images, an adaptive weight block is proposed in our model to determine whether source pixels are focused or not based on the gradient. Under this guidance, we design a special loss function for forcing the fused image to have the same distribution as the focused regions in source images. In addition, a generator and a discriminator are trained to form a stable adversarial relationship. The generator is trained to generate a real-like fused image, which is expected to fool the discriminator. Correspondingly, the discriminator is trained to distinguish the generated fused image from the ground truth. Finally, the fused image is very close to ground truth in probability distribution. Qualitative and quantitative experiments are conducted on publicly available datasets, and the results demonstrate the superiority of our ACGAN over the state-of-the-art, in terms of both visual effect and objective evaluation metrics.

Keywords Multi-focus image fusion · Adaptive weight block · Generative adversarial networks · End-to-end

1 Introduction

Due to the limitations of optical lenses, it is often difficult for an imaging device to take an image in which all the objects are captured in focus [13]. Thus, only the objects within the depth-of-field (DOF) have sharp appearance in the photograph while other objects are likely to be blurred. Multi-focus image fusion is known as a valuable technique to obtain an all-in-focus image by fusing multiple images of the same scene taken with different focal settings, which

is beneficial for human or computer operators, and for further image-processing tasks, e.g., segmentation, feature extraction and object recognition [15, 18]. Therefore, multi-focus image fusion has become a significant research topic in the field of image processing [10].

In the past few decades, many methods for multi-focus image fusion are continually proposed by researchers, and these methods can be attributed to two categories: spatial domain methods and transform domain methods. The methods based on spatial domain can be further divided into three groups according to different fusion rules [2, 9, 10]: pixel-based, block-based, and region-based fusion methods. Among them, the activity level measurements generally adopt the gradient information as a reference. In terms of transform domain methods, after the source images are transferred into other transform domains, the fusion process is mainly implemented in the transformed domains according to the characteristics of the domains. The methods based on transform domain includes discrete wavelet transform (DWT) [30], nonsubsampling contourlet transform (NSCT) [9], sparse representation (SR) [28, 33], subspace [29], etc.

✉ Fan Fan
fanfan@whu.edu.cn

Jun Huang
junhwong@whu.edu.cn

Zhuliang Le
lezhuliang@whu.edu.cn

Yong Ma
mayong@whu.edu.cn

Xiaoguang Mei
meixiaoguang@gmail.com

¹ Electronic Information School, Wuhan University, Wuhan 430072, China

The existing fusion methods present excellent performance in some respects. However, there are still some shortages. First, existing methods often require manual design of activity level measurements and fusion rules, which become more complex and inadequate. Second, generating a decision map is a very common step in the existing multi-focus fusion methods, which is more likely to be a classification problem based on sharpness detection. However, although these methods can correctly classify in most regions, it is often difficult to accurately determine the focused and defocused regions well near the boundary lines.

With the unprecedented success of deep learning, some deep learning-based fusion methods have been proposed. We will discuss the detailed exposition of deep learning-based fusion methods later in Sect. 2.1. These works have provided new ideas for multi-focus image fusion and achieved promising performance. Nevertheless, there are still some aspects need to be improved. On the one hand, the deep learning framework is generally only applied to a small part, e.g., feature processing, while the overall fusion process is still in traditional frameworks. On the other hand, almost all methods based on deep learning face the need for post-processing, such as consistency checks and decision map optimization, which is not end-to-end strictly.

In order to address the above challenges, in this paper, based on deep learning, we propose a novel end-to-end model for multi-focus image fusion, termed as a generative adversarial network with adaptive constraints (ACGAN). Due to the different gradient distribution (clear or blurred) between the corresponding pixels of two source images, direct fusion will result in the fused image between clarity and blur, *i.e.*, neutralization, which is not the result we expect. Therefore, to obtain a clear fused image, an adaptive weight block is employed in our model to determine whether source pixels are focused or not based on the gradient. Concretely, the focused pixel shares a bigger gradient, which is selected to be the input of generator to obtain the fused image. In other words, two score maps are generated for source images, which serve as the reference to our specific loss function. In a result, the generator is forced to generate a fused image that is consistent with the focused regions in source images. In addition, to make the fused image more similar to ground truth, a discriminator network is applied to assess whether the fused image is indistinguishable from the ground truth. In the stable adversarial process between the generator and discriminator, more information, e.g., texture details and spatial information, can be preserved to meet this high-level goal. In general, the advantages of our ACGAN are concluded as follows: First, our method is an end-to-end model without manually designing complex activity level measurements and fusion rules, nor does it need any postprocessing.

Second, our method does not need to generate decision map in the intermediate process, but extracts and reconstructs pixel information into a fused image in pixel units, so there is no blurring near the boundary line. Finally, the adaptive weight block in our method guides the generator to generate a fused image that is consistent with the focused regions in source images, which will not suffer from the neutralization phenomenon.

The major contributions of this paper involve the following three aspects: First, the proposed ACGAN is an end-to-end deep learning-based method, which gets rid of manually designing complex activity level measurements and fusion rules, and does not require any postprocessing. Second, the adaptive weight block based on gradient is proposed in our method, guiding the generator to adaptively learn the distribution of the focused pixels. Third, our fused results have good visual effect, which can avoid the problems of blurring near the boundary line in decision map-based methods and the neutralization phenomenon in non-decision map-based methods.

The remainder of this paper is arranged as follows. Sect. 2 describes some related work, including an overview of existing deep learning-based fusion methods and a theoretical introduction of GANs and LSGAN. In Sect. 3, we introduce our method, *i. e.*, ACGAN, with the problem formulation, loss functions and network architectures. Qualitative and quantitative comparisons and ablation experiments are performed in Sect. 4. We conclude in Sect. 5.

2 Related work

In this section, a brief introduction of the existing deep learning-based image fusion methods is given. Moreover, we also present a brief explanation of generative adversarial networks (GAN) and an improved network, namely LSGAN employed in our work.

2.1 Multi-focus image fusion based on deep learning

The deep learning-based methods are mainly based on convolutional neural networks (CNN) and GAN. In the methods based on CNN, Liu et al. [13] applied the convolutional neural network to the multi-focus image fusion task for the first time, and the CNN is used here to classify focused and defocused regions in order to generating a decision map for fusion. Du et al. [4] regarded the detection of decision map as an image segmentation problem between the focused and defocused regions from source images, and achieved segmentation through a multi-scale convolutional neural network. Ma et al. [15] proposed an

unsupervised encoder-decoder model, termed as SESF-Fuse. In contrast to previous works, SESF-Fuse analysed sharp appearance in deep feature instead of original image. As for the methods based on GAN, Ma et al. [17, 20] adopted GAN to the image fusion task for the first time, which is also an unsupervised framework, named Fusion-GAN. Innovatively, Xu et al. [19, 26] addressed multi-resolution image fusion problem with an additional discriminator, and established two adversarial games between a generator and two discriminators to generate a fused image. Then, Guo et al. [6] proposed FuseGAN for multi-focus image fusion with least square GAN to enhance the training stability. In addition, Xu et al. [27, 32] proposed two frameworks for uniform image fusion, which can address multi-focus, multi-modal and multi-exposure image fusion.

2.2 Least square GAN

The GAN is first proposed by Goodfellow et al. [5] in 2014, which is one of the generative models. The generator G and discriminator D included in the GAN are two adversarial models, where the generative model G captures the data distribution and the discriminative model D is used to determine whether the input is a generated sample or a real sample. In addition, an adversarial game is established between G and D . Particularly, the generator aims to generate a sample to fool the discriminator, while the discriminator tries to determine whether a sample is from the real sample or not. Finally, the sample generated by the generator cannot be distinguished by the discriminator.

In the following years after the advent of GAN, many variants of GAN are proposed [12, 31]. Specifically, in 2017, Mao et al. [22] proposed the least square GAN, *i.e.*, LSGAN, to improve the stability of training process. The sigmoid cross entropy loss function for the discriminator adopted in the regular GAN may lead to the gradient-vanishing problem when training. Therefore, the least squares loss function for the discriminator is introduced in LSGAN to address the above mentioned problem. The optimization functions for LSGAN are shown as follows:

$$\min_D V_{LSGAN}(D) = \frac{1}{2} \mathbb{E}_{x \sim P_{data}(x)} [(D(x) - b)^2] + \frac{1}{2} \mathbb{E}_{z \sim P_z(z)} [(D(G(z)) - c)^2], \tag{1}$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2} \mathbb{E}_{z \sim P_z(z)} [(D(G(z)) - a)^2]. \tag{2}$$

where b and c denote the labels for real data and fake data, respectively, and a is the label that the generator expects the discriminator to believe for fake data. One of the optimization strategies is to set the $b - a = 1$ and $b - c = 2$, which minimizing the χ^2 divergence between $P_{data} + P_g$ and $2P_g$.

The other is to set $a = b$, which can force the generated samples to be more similar to the real ones.

3 Proposed method

In this section, with analysis of the characteristics of multi-focus images, we provide our problem formulation with the proposed adaptive weight block, the definition and design of loss functions. At the end of this section, we present the design of network architecture concretely.

3.1 Problem formulation

Multi-focus images are images with different focused regions. The essence of multi-focus image fusion is to extract and integrate the most important information in the source images, *i.e.*, the focused regions, to a single image. The focused region can be characterized by the intensity distribution and texture details. The entire fusion procedure is shown in Fig. 1.

To extract and integrate the focused regions in source images, we propose an adaptive weight block, which is employed to evaluate the sharpness of each pixel based on the gradient, as presented in Fig. 2. The focused regions share bigger gradient. Specifically, the pixels with larger gradient are selected by us as the optimization target at the corresponding pixel positions of the two source images, while the smaller ones are abandoned. Therefore, the specific content loss function designed by us with the adaptive weight block can adaptively guide the fused image to approximate the intensity distribution and gradient distribution of the focused regions from source images

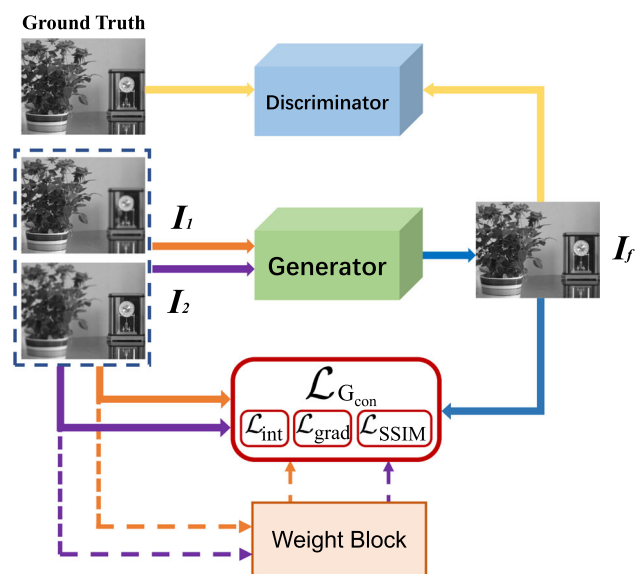


Fig. 1 The fusion procedure of the proposed ACGAN



Fig. 2 The source images and corresponding gradient maps. From left to right: source image 1, source image 2, the gradient map of source image 1, and the gradient map of source image 2

at the pixel level. The ablation experiment of the adaptive weight block is also conducted later in Sect. 4.5.1. In addition, since the goal of our optimization is based on each pixel, in order to avoid chromatic aberrations in the fused image and ensure the overall naturalness of it, we add the SSIM loss term. Based on the principle of statistics, the mean of the larger scores in each source image patch is calculated as the weight of corresponding SSIM loss term. The effect of SSIM loss term will be verified later in Sect. 4.5.2. Working with the adaptive weight block on content loss, our ACGAN can simultaneously achieve the clear and natural fused image.

To further improve the quality of the fused image and bring it closer to our ideal ground truth, we add a discriminator to establish an adversarial relationship with the generator. The generator aims to generate a real-like image based on our specifically designed content loss to fool its corresponding discriminator, while the discriminator aims to distinguish the differences between the generated image and ground truth. Finally, the discriminator cannot distinguish the generated image from ground truth, and the fused image can reach a higher quality, *i. e.*, richer texture details and more spatial information. The influence of the additional discriminator will be analyzed later in Sect. 4.5.3.

3.2 Loss function

The loss function in our work can be divided into the loss of generator \mathcal{L}_G and the loss of discriminator \mathcal{L}_D .

3.2.1 Generator loss

The loss function of generator \mathcal{L}_G consists of content loss $\mathcal{L}_{G_{con}}$ and adversarial loss $\mathcal{L}_{G_{adv}}$. Due to the instability of GAN, the introduction of content loss adds a series of constraints to the generator to achieve the fusion goal, while the adversarial loss allows the fused image to meet stricter requirements. \mathcal{L}_G is defined as follows:

$$\mathcal{L}_G = \mathcal{L}_{G_{con}} + \mathcal{L}_{G_{adv}}. \tag{3}$$

Among them, $\mathcal{L}_{G_{con}}$ includes intensity loss, gradient loss and SSIM loss, which can be expressed as follows:

$$\mathcal{L}_{G_{con}} = \mathcal{L}_{int} + \alpha_1 \mathcal{L}_{grad} + \alpha_2 \mathcal{L}_{SSIM}, \tag{4}$$

where α_1 and α_2 are used to control the trade-off, which will be analyzed later in Sect. 4.5.4.

The adaptive weight block acts on the intensity loss \mathcal{L}_{int} and gradient loss \mathcal{L}_{grad} , which guides the generator to generate a fused image that is consistent with the focused regions in pixel level. Concretely, the \mathcal{L}_{int} can guide the fused image to have the same intensity distribution as the focused regions in source images, which is presented as follows:

$$\begin{aligned} \mathcal{L}_{int} = & \frac{1}{HW} \sum_i \sum_j (S_{1_{ij}} - \min(S_{1_{ij}}, S_{2_{ij}})) \\ & \cdot (I_{f_{ij}} - I_{1_{ij}})^2 + (S_{2_{ij}} - \min(S_{1_{ij}}, S_{2_{ij}})) \\ & \cdot (I_{f_{ij}} - I_{2_{ij}})^2, \end{aligned} \tag{5}$$

where H and W mean the height and width of the source images, *i. e.*, I_1 and I_2 , and the fused image, *i. e.*, I_f . In particular, $S_{(\cdot)}$ is the score map generated by the adaptive weight block based on the gradient, whose size is also $H \times W$. i and j mean the pixel in the i -th row and the j -th column. The $\min(S_{1_{ij}}, S_{2_{ij}})$ means the minimum gradient score of the corresponding pixel in the source images.

Similarly, the \mathcal{L}_{grad} is employed to guide the fused image to have the same gradient distribution, *i. e.*, texture details, as the focused regions in source images. \mathcal{L}_{grad} is formalized as follows:

$$\begin{aligned} \mathcal{L}_{grad} = & \frac{1}{HW} \sum_i \sum_j (S_{1_{ij}} - \min(S_{1_{ij}}, S_{2_{ij}})) \\ & \cdot (\nabla I_{f_{ij}} - \nabla I_{1_{ij}})^2 + (S_{2_{ij}} - \min(S_{1_{ij}}, S_{2_{ij}})) \\ & \cdot (\nabla I_{f_{ij}} - \nabla I_{2_{ij}})^2. \end{aligned} \tag{6}$$

On this basis, we employ the SSIM loss term to avoid chromatic aberrations in the fused image and ensure the overall naturalness of it. It is worth noting that, for each overall source image, structural information with a larger average gradient is preserved. Specifically, the \mathcal{L}_{SSIM} is defined as follows:

$$\begin{aligned} \mathcal{L}_{SSIM} = & \frac{1}{HW} \sum_i \sum_j (S_{1_{ij}} - \min(S_{1_{ij}}, S_{2_{ij}})) \\ & \cdot (1 - SSIM_{I_f, I_1}) + (S_{2_{ij}} - \min(S_{1_{ij}}, S_{2_{ij}})) \\ & \cdot (1 - SSIM_{I_f, I_2}), \end{aligned} \tag{7}$$

where SSIM stands for structural similarity and is an indicator for measuring the similarity between the source images and the fused image. The larger the SSIM, the more similar the structure of the fused image is to the source image. Mathematically, SSIM is defined as follows:

$$\begin{aligned} SSIM_{X,F} = & \sum_{x,f} \frac{2\mu_x\mu_f + C_1}{\mu_x^2 + \mu_f^2 + C_1} \cdot \frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2 + \sigma_f^2 + C_2} \\ & \cdot \frac{\sigma_{xf} + C_3}{\sigma_x\sigma_f + C_3}, \end{aligned} \tag{8}$$

where the three items on the right hand reflect the comparisons of brightness, contrast and structural, respectively. x and f express the image patches in source image X and fused image F . μ denotes the mean value, while σ denotes the standard deviation/covariance.

The adversarial loss of the generator $\mathcal{L}_{G_{adv}}$ is used to force the fused image to achieve a higher quality, which is formalized as follows:

$$\mathcal{L}_{G_{adv}} = \frac{1}{N} \sum_{n=1}^N (D(I_f^n) - a)^2, \tag{9}$$

where N denotes the number of fused image, and we employ a as the probability label that the generator expects the discriminator to judge the fused image.

3.2.2 Discriminator loss

The discriminator in ACGAN plays a role of discriminating between the ground truth and the generated fused image. The adversarial loss of discriminator can calculate the least square loss to identify whether the distribution in fused image is unrealistic, and encourage the fused image to match the realistic distribution. The discriminator loss \mathcal{L}_D is defined as follows:

$$\mathcal{L}_D = \frac{1}{N} \sum_{n=1}^N [D(I_f^n) - b]^2 + [D(I_g^n) - c]^2, \tag{10}$$

where b is the random label of the fused image, which is expected to be small enough, while c is the random label of ground truth, which is expected to be large enough, as the fused image is expected to be judged by discriminator as fake data, while the ground truth is expected to be real data.

3.3 Network architecture

3.3.1 Generator architecture

The network architecture of generator is illustrated in Fig. 3. The design of our generator draws on the idea of the pseudo-siamese network. For two different source images, we use different parameters to extract different features with two branches, which is suitable for processing source images with different focused regions. Adequate information exchange is the biggest characteristics in our generator, which is reflected in the following three parts.

First, the information exchange on each branch as shown in red, green and purple arrows: Similar to DenseNet [22], each layer is established a short direct connection with other layers in a feed-forward fashion. Avoiding vanishing gradients, strengthening feature propagation and reducing the number of parameters are the main advantages of this design. In particular, the convolution kernel of the first convolutional layer is 5×5 , while the others in the next three convolutional layers are 3×3 . Second, the information exchange between branches as shown in blue arrows: The information between branches is also exchanged by concatenating and convolution, which can be seen as “pre-fusion”. Third, the final fusion: the outputs of two branches are concatenating together, which is the input of the last convolutional layer. The output of the last convolutional layer with the kernel size of 1×1 is the fused image. It is worth noting that throughout the process we use “SAME” as the padding mode to keep the size of the feature map consistent with source images.

3.3.2 Discriminator architecture

The discriminator is designed to establish an adversarial relationship with the generator. Particularly, it aims to distinguish the generated images from the ground truth, which is illustrated in Fig. 4. There are four convolution layers with the kernel size of 3×3 and one linear layer with the kernel size of 1×1 in the discriminator. The leaky ReLU activation function is employed in all four convolution layers with the stride of 2. We use the last linear layer to acquire the probability scalar.

4 Experimental results and analysis

In this section, we validate the effectiveness of our ACGAN by comparing it with several state-of-the-art methods on publicly available datasets. Not only the qualitative comparisons but also the quantitative comparisons are implemented in our work. We also conduct the

Fig. 3 The network architecture of generator

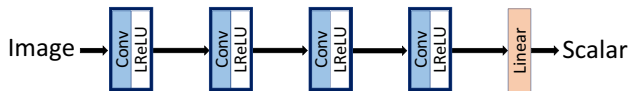
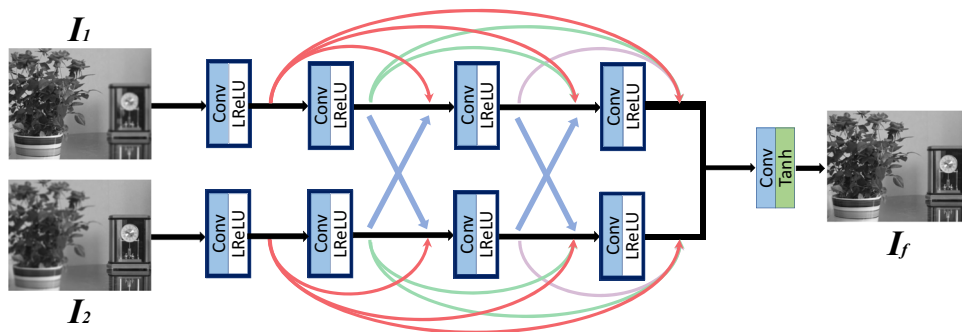


Fig. 4 The network architecture of discriminator

ablation experiments of the adaptive weight block, the SSIM loss term, and the discriminator. Moreover, the analysis of α_1 and α_2 is also performed.

Algorithm 1: Training procedure of ACGAN

```

1: for  $m$  epochs do
2:   for  $n$  steps do
3:     for  $x$  times do
4:       Select  $k$  patches from source 1  $\{I_1^1, I_1^2 \dots I_1^k\}$ ;
5:       Select  $k$  patches from source 2  $\{I_2^1, I_2^2 \dots I_2^k\}$ ;
6:       Select  $k$  fused patches  $\{I_f^1, I_f^2 \dots I_f^k\}$ ;
7:       Update discriminator parameters  $\theta_D$  by AdamOptimizer to
         minimize  $\mathcal{L}_D$  in Eq. (10)
8:     end for
9:     Select  $k$  patches from source 1  $\{I_1^1, I_1^2 \dots I_1^k\}$ ;
10:    Select  $k$  patches from source 2  $\{I_2^1, I_2^2 \dots I_2^k\}$ ;
11:    Update generator parameters  $\theta_G$  by AdamOptimizer to minimize
       $\mathcal{L}_G$  in Eq. (3)
12:   end for
13: end for
    
```

The detailed training procedure is summarized in Alg. 1. We train the generator and discriminator iteratively to establish an adversarial relationship. Among them, our total number of epoch is m , it takes n steps to train each epoch, the number of training generator is x times of the number of training discriminator, and the batch size is set as k . Concretely, m , x and k is set to 20, 2 and 32, respectively. The n is the ratio between the whole number of patches and batch size. We update all parameters by AdamOptimizer in our ACGAN. Moreover, we set $\alpha_1 = 3$ and $\alpha_2 = 10$ in Eq. (4).

In addition, the images in training data are grayscale images with single channel, while the images in testing data are color images with RGB channels. In order to fuse the images in testing data with the trained model, $YCbCr$ color space is employed in our work. Y channel (luminance channel) can represent structural details and the brightness variation, which is devoted to participating in fusion. Cb and Cr channels are chrominance channels, which should not be changed. Finally, the fused image is transferred back to RGB color space with Cb and Cr channels to acquire the final result.

4.1 Experimental settings

The dataset we train our network is from a public dataset website¹. In order to verify the generalization ability of our model, we test our network in different public multi-focus image datasets, *i.e.*, Lytro dataset [23] and some standard images for multi-focus image fusion². The image pairs have been accurately aligned, and image registration techniques are required for unaligned images [16, 21]. When training, the expansion strategy of tailoring and decomposition is employed in our work to get a larger data set, and the training set is cropped to 23, 714 groups of size 60×60 with two source images and one ground truth in each group. We employ 30 image pairs from the two datasets for testing.

¹ <https://sites.google.com/view/durgaprasadbavirisetti/datasets>.

² <https://www.mathworks.com/matlabcentral/fileexchange/45992-standard-images-for-multifocus-image-fusion>.

4.2 Comparative methods and evaluation metrics

We select five state-of-the-art methods to evaluate our ACGAN on publicly available datasets, including, GFDF [24], DSIFT [14], S-A [11], CNN [13] and SESF [15]. In order to have a comprehensive assessment. CNN and SESF are methods based on deep learning, while others are traditional methods, and GFDF, DSIFT, CNN and SESF are methods based on the decision map.

In order to have a more accurate evaluation of the experimental results. we utilize six metrics to evaluate the fusion results, including, sum of the correlations of differences (SCD) [1], visual information fidelity (VIF) [8], correlation coefficient (CC) [3], $Q^{AB/F}$, which measure between fused image and source images, and entropy (EN) [25], standard deviation (SD) [25], which measure the fused image itself.

4.3 Qualitative comparisons

The intuitive results on four typical image pairs are shown in Fig. 5. Our ACGAN not only performs well on the overall image but also in local details, especially for the boundary of focused and defocused regions. As can be seen in the enlarged regions in the red boxes in the above two groups of results, the results of GFDF, DSIFT, CNN and SESF that are all based on decision map cannot accurately retain details near the junction of focused and defocused regions, and lose details due to misclassification, e.g., the pip on the ceiling and details between fingers. On the contrary, our ACGAN can accurately preserve the details in the focused regions. In addition, as for the remaining comparative method that are not based on the decision map, such as S-A, it suffers from the neutralization phenomenon and blurring near the boundary line, e.g., the details between the fingers in the upper right group, the edge of the hat in the bottom left group and the building behind the monkey in the bottom right group. By comparison, our ACGAN can preserve them better.

4.4 Quantitative comparisons

The quantitative comparisons of our ACGAN with the competitors on the 30 image pairs in the dataset are also reported, which is summarized in Fig. 6. As can be seen from the statistical results, our ACGAN can achieve the largest mean values on all six metrics. These results

demonstrate that our method has the greatest correlation with source images and the best contrast, and the edge information can be preserved to the greatest extent. In addition, our method can perform the best visual effect.

In order to verify the convenience of our method, the mean and standard deviation of running time for our ACGAN and the competitors are presented in Table 1, where the methods, *i.e.*, SESF and our ACGAN are tested on GPU RTX 2080Ti, while other methods are tested on CPU i7-8750H (The testing environments of the competitors are consistent with the original paper). Clearly, our ACGAN can also perform comparable efficiency.

4.5 Ablation experiments

4.5.1 Adaptive weight block analysis

The adaptive weight block is employed in our model to guide the generator to adaptively learn the distribution of the focused pixels, avoiding the neutralization phenomenon. In order to show the effect of the adaptive weight block, we perform the following comparative experiments: (a) The adaptive weight block is not employed. (b) The adaptive weight block is employed. The experimental settings of two comparative experiments are the same and the results are shown in the Fig. 7. By comparison, The fused result without the adaptive weight block suffers from the neutralization phenomenon, while the fused result with the adaptive weight block can present

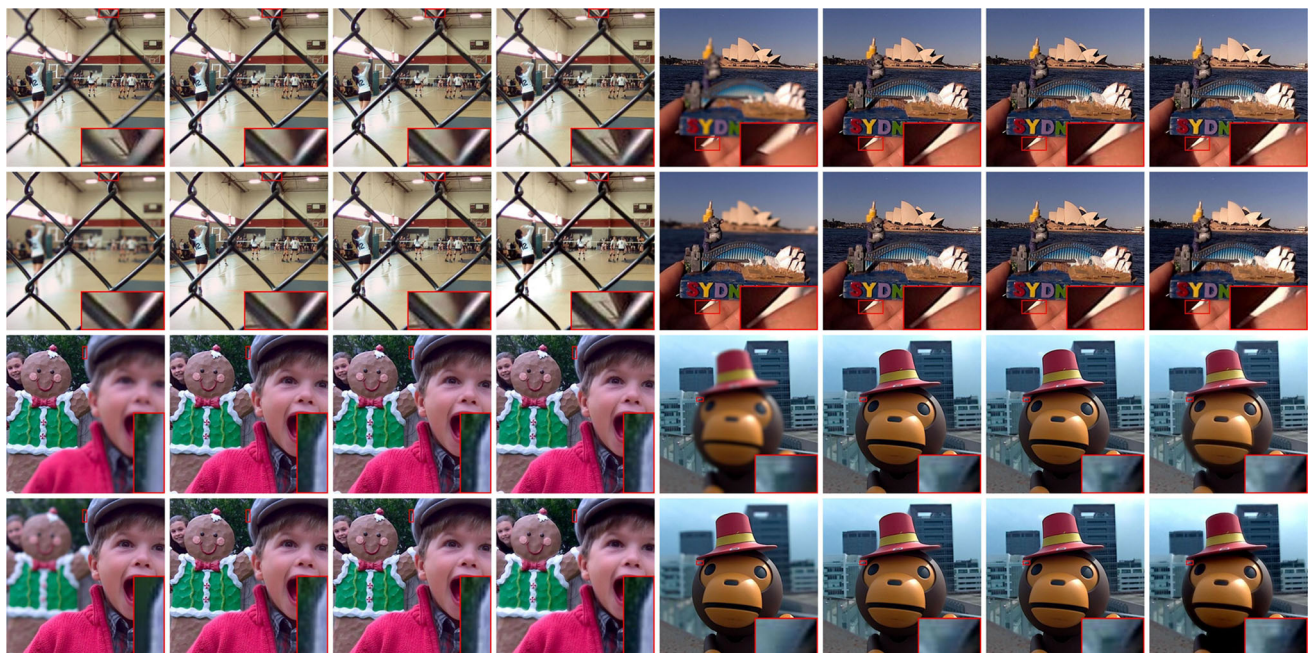


Fig. 5 Qualitative results on the Lytro dataset. In each group, the first column are the source images; the second column are the results of GFDF (2019) and CNN (2017); the third column are the results of

DSIFT (2015) and SESF (2019), and the fourth column are the results of S-A (2018) and our ACGAN

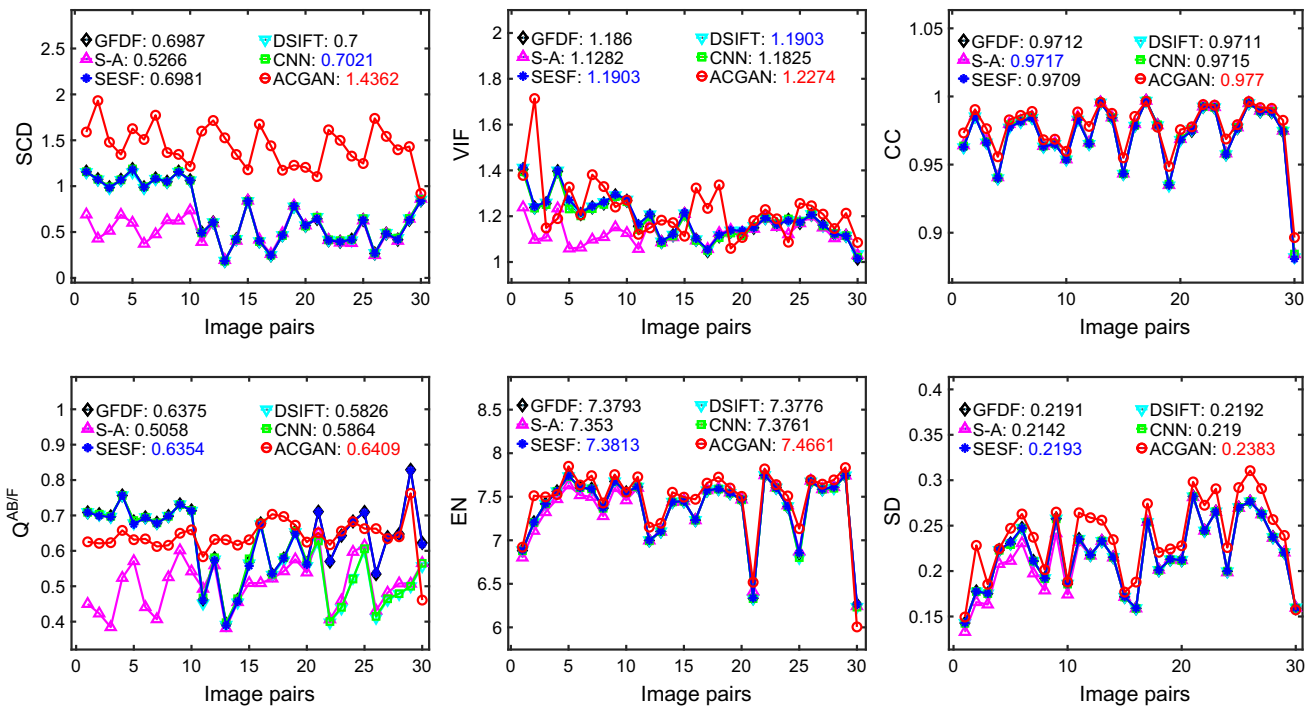


Fig. 6 Quantitative comparison of our ACGAN with five state-of-the-art methods. Means of metrics for different methods are shown in the legends. Optimal values are shown in red and suboptimal values in blue

Table 1 The mean and standard deviation of running time in different methods. (unit: second)

Methods	GFDF [7]	DSIFT [14]	S-A [11]	CNN [13]	SESF [15]	ACGAN
Mean	0.2816	5.8540	0.2435	116.9590	0.3396	0.0421
STD	0.1251	3.5241	0.1032	54.3958	0.2245	0.0221

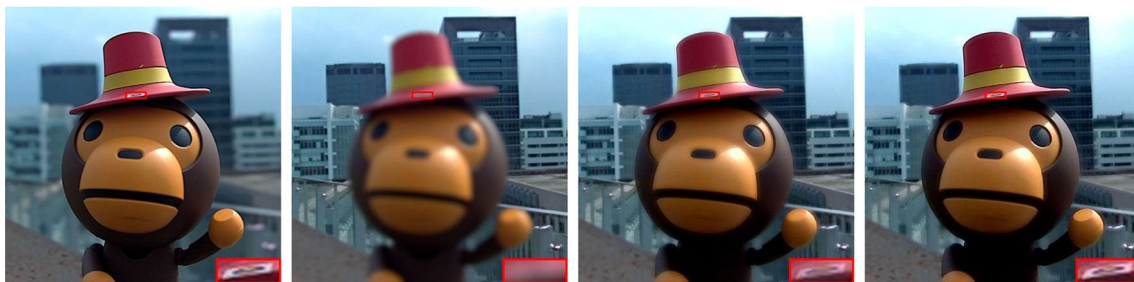


Fig. 7 Ablation experiment of the adaptive weight block. From left to right: source image 1, source image 2, the fused result without adaptive weight block and the result with adaptive weight block

the focused regions well. As a result, it proves that the adaptive weight block can avoid the neutralization phenomenon well.

4.5.2 SSIM loss term analysis

In order to make the fused image more similar to the focused regions in the source image, including the color distribution and overall naturalness, the SSIM loss term is introduced to address the above issues. The effect of the

SSIM loss term is verified by the following comparative experiments: (c) The SSIM loss term is not employed. (d) The SSIM loss term is employed. The experimental settings of two comparative experiments are the same and the results are shown in the Fig. 8. The fused image with the SSIM loss term employed has almost the same color distribution as the focused area in the source image. On the other hand, the fused image without the SSIM loss term suffers from the chromatic aberrations with darker color, whose overall naturalness is also worse than the other one.



Fig. 8 Ablation experiment of the SSIM loss term. From left to right: source image 1, source image 2, the fused result without SSIM loss term and the result with SSIM loss term



Fig. 9 Ablation experiment of the discriminator. From left to right: source image 1, source image 2, the fused result without discriminator and the result with discriminator

Therefore, it can be concluded that the SSIM loss term has a positive impact on the fused image.

4.5.3 Discriminator analysis

We use the discriminator to establish a stable adversarial relationship with the generator, forcing the fused image to be more similar to ground truth, *i.e.*, the focused regions in source images. In order to show the effect of the discriminator. The following comparative experiments are performed: (e) The discriminator is not employed. (f) The discriminator is employed. The experimental settings of two comparative experiments are the same and the results are shown in the Fig. 9. The result with the discriminator is more similar to the focused regions in the source image. In contrast, the result without the discriminator suffers from more blurred details. It can be seen that the discriminator plays an important role in the fusion process.

4.5.4 Parameter analysis

In our work, \mathcal{L}_{int} and $\mathcal{L}_{\text{grad}}$ are employed to guide the fused image to have the same intensity and gradient distribution as the focused regions in source images, and the $\mathcal{L}_{\text{SSIM}}$ is used to avoid chromatic aberrations in the fused image and ensure the overall naturalness of it based on \mathcal{L}_{int} and $\mathcal{L}_{\text{grad}}$. Therefore, in order to obtain the optimal values of α_1 and α_2 , we first analyze α_1 without $\mathcal{L}_{\text{SSIM}}$. We select 5 values

(0.3, 1.5, 3, 4.5 and 6) for α_1 , and determine the optimal value of α_1 by comparing the results of quantitative comparison, which is summarized in Fig. 10. As can be seen from the statistical results, when $\alpha_1 = 3$, the results of the quantitative comparison are optimal overall. Therefore, parameter α_1 is determined to be set to 3.

Next, based on $\alpha_1 = 3$, we add the $\mathcal{L}_{\text{SSIM}}$ loss term for a higher fusion quality. Similarly, we also select 5 values (1, 5, 10, 15 and 20) for α_2 , and determine the optimal value of α_2 by comparing the quantitative comparison results, which is summarized in Fig. 11. As can be seen from the statistical results, when $\alpha_2 = 10$, the results of the quantitative comparison are optimal overall. Therefore, parameter α_2 is determined to be set to 10.

5 Conclusion and future work

In this paper, we propose a new end-to-end model for multi-focus image fusion based on generative adversarial networks, termed as ACGAN. Our ACGAN overcomes the difficulty of neutralization phenomenon and blurring near the boundary line with an adaptive weight block. In addition, an adversarial relationship between the generator and discriminator is established to generate the fused images of higher quality. For qualitative experiments, our ACGAN not only performs well on the overall image but also in local details, especially for the boundary of focused and

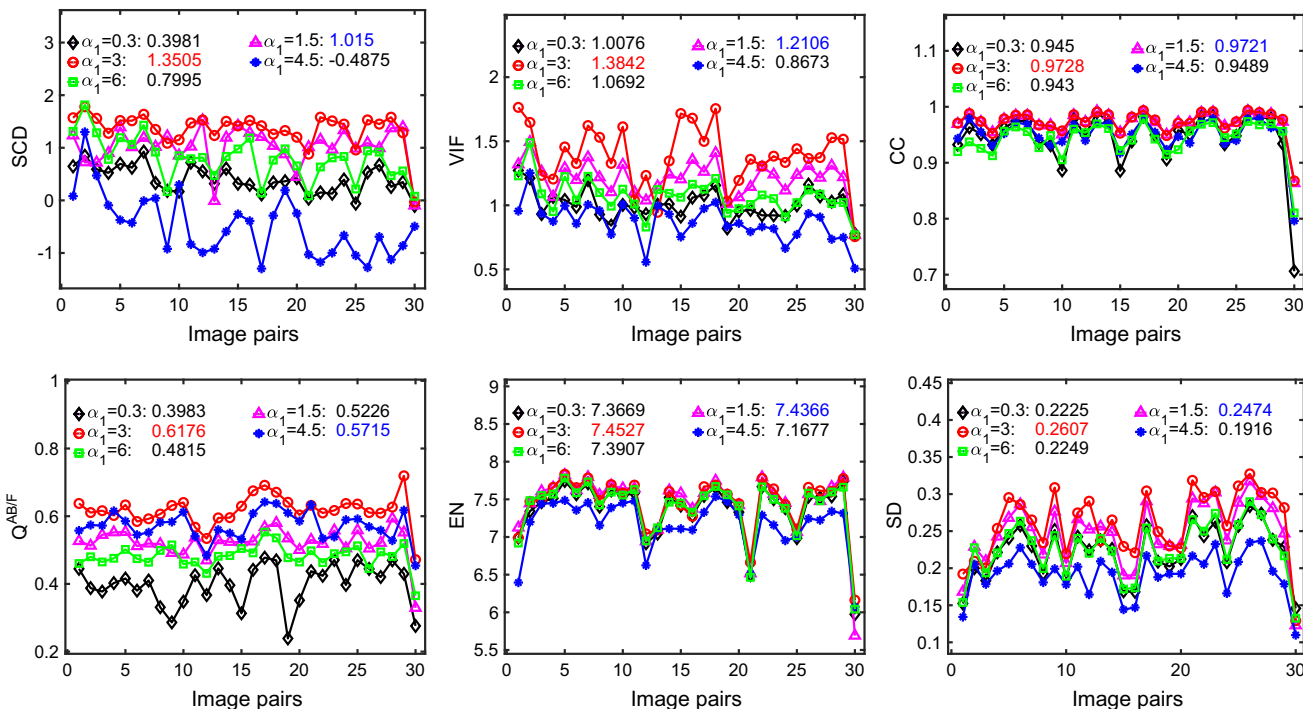


Fig. 10 Quantitative comparison of different α_1 values. Means of metrics for different α_1 values are shown in the legends. Optimal values are shown in red and suboptimal values in blue (color figure online)

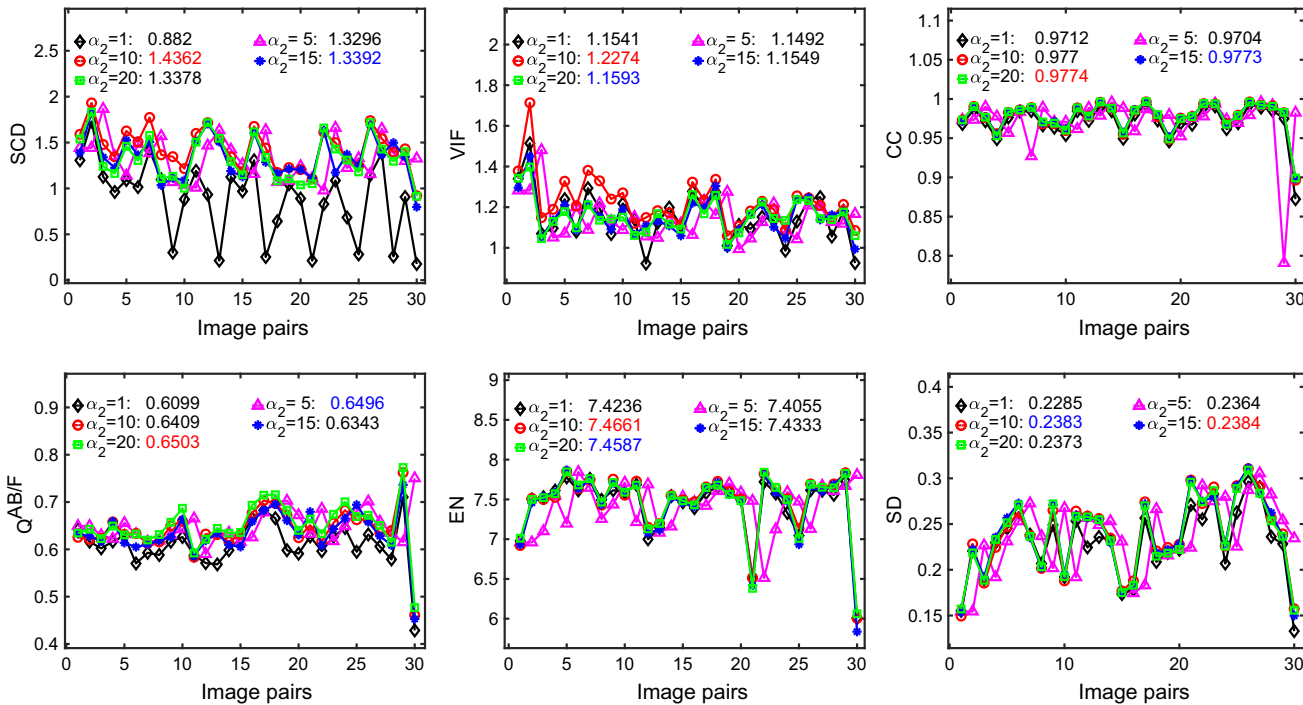


Fig. 11 Quantitative comparison of different α_2 values. Means of metrics for different α_2 values are shown in the legends. Optimal values are shown in red and suboptimal values in blue (color figure online)

defocused regions. Quantitative experiments verify that our method performs better than the existing state-of-the-art methods on six widely used metrics.

There may be potential limitation in our work, and our method is not based on decision map. In the existing methods based on decision map, the pixels of the fused

image are completely consistent with the pixels of the source images. In contrast, the pixels in our fused image are obtained by learning the pixels in the focused regions in the source images. Although it can overcome the problem of blurring near the boundary line in the existing decision map-based methods and present good visual effect, it is difficult for the pixels in our fused image to be completely the same as the pixels in the focused regions in the source images. Therefore, in our future work, we will be committed to solving the problem of blurring near the boundary line based on decision map.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 61903279.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Aslantas V, Bendes E (2015) A new image quality metric for image fusion: the sum of the correlations of differences. *AEU Int J Electron Commun* 69(12):1890–1896
- Chen J, Li X, Luo L, Mei X, Ma J (2020) Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf Sci* 508:64–78
- Deshmukh M, Bhosale U (2010) Image fusion and image quality assessment of fused images. *Int J Image Process* 4(5):484
- Du C, Gao S (2017) Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network. *IEEE Access* 5:15750–15761
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Advances in neural information processing systems*, pp 2672–2680
- Guo X, Nie R, Cao J, Zhou D, Mei L, He K (2019) Fusegan: learning to fuse multi-focus image via conditional generative adversarial network. *IEEE Trans Multimed* 21:1982–1996
- Haghighat MBA, Aghagolzadeh A, Seyedarabi H (2011) Multi-focus image fusion for visual sensor networks in DCT domain. *Comput Electr Eng* 37(5):789–797
- Han Y, Cai Y, Cao Y, Xu X (2013) A new image fusion performance metric based on visual information fidelity. *Inf Fusion* 14(2):127–135
- Li H, Chai Y, Li Z (2013) Multi-focus image fusion based on nonsubsampling contourlet transform and focused regions detection. *Optik Int J Light Electron Opt* 124(1):40–51
- Li S, Kang X, Hu J, Yang B (2013) Image matting for fusion of multi-focus images in dynamic scenes. *Inf Fusion* 14(2):147–162
- Li W, Xie Y, Zhou H, Han Y, Zhan K (2018) Structure-aware image fusion. *Optik* 172:1–11
- Liu L, Zhang H, Xu X, Zhang Z, Yan S (2019) Collocating clothes with generative adversarial networks cosupervised by categories and attributes: a multidiscriminator framework. *IEEE Trans Neural Netw Learn Syst*. <https://doi.org/10.1109/TNNLS.2019.2944979>
- Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. *Inf Fusion* 36:191–207
- Liu Y, Liu S, Wang Z (2015) Multi-focus image fusion with dense sift. *Inf Fusion* 23:139–155
- Ma B, Ban X, Huang H, Zhu Y (2019) Sef-fuse: An unsupervised deep model for multi-focus image fusion. *arXiv preprint arXiv:1908.01703*
- Ma J, Jiang X, Jiang J, Zhao J, Guo X (2019) LMR: learning a two-class classifier for mismatch removal. *IEEE Trans Image Process* 28(8):4045–4059
- Ma J, Liang P, Yu W, Chen C, Guo X, Wu J, Jiang J (2020) Infrared and visible image fusion via detail preserving adversarial learning. *Inf Fusion* 54:85–98
- Ma J, Ma Y, Li C (2019) Infrared and visible image fusion methods and applications: a survey. *Inf Fusion* 45:153–178
- Ma J, Xu H, Jiang J, Mei X, Zhang XP (2020) DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans Image Process* 29:4980–4995
- Ma J, Yu W, Liang P, Li C, Jiang J (2019) FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inf Fusion* 48:11–26
- Ma J, Zhao J, Jiang J, Zhou H, Guo X (2019) Locality preserving matching. *Int J Comput Vis* 127(5):512–531
- Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*, pp 2794–2802
- Nejati M, Samavi S, Shirani S (2015) Multi-focus image fusion using dictionary-based sparse representation. *Inf Fus* 25:72–84
- Qiu X, Li M, Zhang L, Yuan X (2019) Guided filter-based multi-focus image fusion through focus region detection. *Signal Process Image Commun* 72:35–46
- Roberts JW, Van Aardt JA, Ahmed FB (2008) Assessment of image fusion procedures using entropy, image quality, and multi-spectral classification. *J Appl Remote Sens* 2(1):023522
- Xu H, Liang P, Yu W, Jiang J, Ma J (2019) Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In: *Proceedings of twenty-eighth international joint conference on artificial intelligence (IJCAI-19)*, pp 3954–3960
- Xu H, Ma J, Le Z, Jiang J, Guo X (2020) FusionDn: A unified densely connected network for image fusion. In: *Proceedings of the thirty-fourth AAAI conference on artificial intelligence*
- Yang B, Li S (2009) Multifocus image fusion and restoration with sparse representation. *IEEE Trans Instrum Meas* 59(4):884–892
- Yang L, Guo B, Ni W (2007) Multifocus image fusion algorithm based on contourlet decomposition and region statistics. In: *Fourth international conference on image and graphics (ICIG 2007)*, pp 707–712. *IEEE*
- Yang Y, Huang S, Gao J, Qian Z (2014) Multi-focus image fusion using an effective discrete wavelet transform based algorithm. *Meas Sci Rev* 14(2):102–108
- Zhang H, Sun Y, Liu L, Wang X, Li L, Liu W (2018) Clothing-out: a category-supervised GAN model for clothing segmentation and retrieval. *Neural Comput Appl*. <https://doi.org/10.1007/s00521-018-3691-y>
- Zhang H, Xu H, Xiao Y, Guo X, Ma J (2020) Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: *Proceedings of the thirty-fourth AAAI conference on artificial intelligence*
- Zhang Q, Liu Y, Blum RS, Han J, Tao D (2018) Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Inf Fus* 40:57–75

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.